Brasil 6G

Projeto Brasil 6G
Fase III

Atividade 3.1 - Aplicações de Visão Computacional para Verticais Estratégicas - Parte 1

15/07/2025

Histórico de Atualizações:

Versão	Data	Autor(es)	Notas
1	15/07/2025	Aldebaro Klautau – UFPA Cristiano Bonato Both – UNISINOS Daniel Takashi N. Suzuki – UFPA Felipe A. P. de Figueiredo – INATEL Matheus Ferreira Silva – INATEL João Borges – UFPA José Ferreira de Rezende – UFRJ Wesley L. Passos – UFRJ	Elaboração de conteúdo
2	15/07/2025	Aldebaro Klautau – UFPA Luciano Leonel Mendes – Inatel Vanessa Mendes Rennó – Inatel	Revisão de texto

Lista de Figuras

Lista de Tabelas

Acrônimos

3GPP *3rd Generation Partnership Project*

5G Quinta Geração de Rede Móvel Celular

5G-NR 5G-*New Radio*

6G Sexta Geração de Rede Móvel Celular

CLIP *Contrastive Language-Image Pre-training*

CSI *Channel State Information*

CSI-RS *Channel State Information Reference Signal*

F-DL *Fusion-based Deep Learning*

GPS *Global Positioning System*

IA Inteligência Artificial

ISAC *Integrated Sensing and Communications*

KDE *Kernel Density Estimation*

LiDAR *Light Detection and Ranging*

LoS *Line-of-Sight*

MEC *Mobile Edge Computing*

MIMO *Multiple Input Multiple Output*

mmWave Ondas Milimétricas

NAS *Neural Architecture Search*

NLoS *Non-Line-of-Sight*

RSRP *Reference Signal Received Power*

SSB *Synchronization Signal Block*

SUMO *Simulation of Urban Mobility*

Tbps Terabits por segundo

THz Terahertz

TPE *Tree-structured Parzen Estimator*

V2I *Vehicle-to-Infrastructure*

YOLO *You Only Look Once*

Sumário

1 Introdução

A Atividade 3.1 da Fase III do Projeto Brasil 6G tem como foco a exploração de aplicações baseadas em visão computacional e Inteligência Artificial (IA) em verticais estratégicas, diante da crescente relevância dessas tecnologias para as redes móveis de próxima geração. O uso de imagens capturadas em tempo real e processadas por algoritmos inteligentes é apontado como uma das principais tendências para o avanço das telecomunicações. Especificamente, esta atividade concentra-se na pesquisa de métodos que utilizam visão computacional e IA para aprimorar o desempenho da Sexta Geração de Rede Móvel Celular (6G), com ênfase especial em técnicas de *beam management*.

O advento das comunicações móveis da Quinta Geração de Rede Móvel Celular (5G) e a evolução rumo ao 6G têm impulsionado o uso de sinais em frequências em Ondas Milimétricas (mmWave), para os quais há grande largura de banda disponível e, consequentemente, permitem altas taxas de transmissão de dados [2]. No entanto, essas altas frequências enfrentam desafios significativos, como forte atenuação, bloqueios frequentes e alcance reduzido. Nesse contexto, o gerenciamento eficiente dos feixes, o *beam management*, tornou-se uma função essencial para manter conexões confiáveis e de alto desempenho em redes 5G e 6G operando em mmWave [3].

Tradicionalmente, as técnicas de *beam management* em redes celulares se baseiam em métodos determinísticos e heurísticos, definidos em especificações como as das Releases 15 e 16 do *3rd Generation Partnership Project* (3GPP). Esses métodos incluem varredura exaustiva de feixes, *feedback* periódico do terminal e algoritmos baseados em medidas de potência do sinal recebido (do inglês, *Reference Signal Received Power* (RSRP)). Embora eficazes, esses procedimentos impõem alta sobrecarga ao sistema, consomem tempo e não se adaptam bem a cenários dinâmicos com mobilidade elevada ou bloqueios imprevisíveis.

Com o avanço das técnicas de IA, novas abordagens baseadas em aprendizado de máquina vêm sendo propostas para aprimorar o gerenciamento de feixes. O 3GPP reconheceu esse potencial e, a partir da Release 18, passou a considerar explicitamente a adoção de IA em três casos de uso principais: *beam management*, estimativa de canal e localização. Em particular, para *beam management*, redes neurais e algoritmos de aprendizado por reforço têm sido bastante explorados para prever feixes ótimos com base em informações contextuais, oriundas de sensores extras como o *Light Detection and Ranging* (LiDAR), reduzindo a necessidade de varredura ativa e aumentando a robustez da conexão [1, 4].

Esta atividade está alinhada com esse esforço internacional ao investigar o uso de visão computacional e técnicas baseadas em IA para melhorar o *beam management* em redes com suporte a mmWave. A proposta é analisar e comparar abordagens convencionais e baseadas em aprendizado de máquina em diferentes cenários, avaliando métricas como tempo de descoberta de feixe, robustez à mobilidade e eficiência espectral. Além disso, o projeto considera a conformidade com as diretrizes e propostas técnicas discutidas nos grupos de trabalho do 3GPP, visando contribuir com soluções práticas e alinhadas aos padrões em evolução.

Para tal, neste relatório o Capítulo 2 desenvolve o método *baseline* para seleção de feixes com base em sinais pilotos, estabelecendo uma referência de desempenho para as abordagens subsequentes. O Capítulo 3 expande bases de dados de canais, com o objetivo de viabilizar simulações mais realistas e compatíveis com aplicações de visão computacional. No Capítulo 4, é explorada a busca de arquiteturas neurais otimizadas para a tarefa de seleção de feixes. O Capítulo 5 apresenta o uso de visão computacional e dados multimodais, avaliando seu potencial para melhorar a seleção de feixes em cenários complexos e dinâmicos. Por fim, Capítulo 6 presenta as conclusões gerais do relatório.

Ao investigar esse tema, o projeto busca não apenas demonstrar ganhos práticos no desempenho das redes, mas também entender melhor os limites e desafios do uso de IA em aplicações de tempo real nas telecomunicações. A expectativa é que os resultados contribuam para as redes móveis 6G, fornecendo subsídios técnicos para novas Releases do 3GPP e para a implementação de sistemas inteligentes, adaptativos e energeticamente eficientes em ambientes urbanos densos e dinâmicos.

2 Método *Baseline* para Seleção de Feixes Usando Pilotos

O desenvolvimento de algoritmos baseados em visão computacional e IA tem ganhado destaque no contexto das redes móveis de próxima geração. No entanto, para que os benefícios dessas abordagens sejam avaliados de forma objetiva, é fundamental que sejam comparadas com métodos tradicionais. Nesse sentido, este capítulo descreve o estabelecimento de um método *baseline*, que serve como referência para análise comparativa com soluções baseadas em IA e visão computacional desenvolvidas ao longo do projeto.

2.1 Definição do problema e escopo da solução

Para estabelecimento de referencial comparativo em relação ao desempenho de métodos propostos no projeto, foi desenvolvido um código para executar o processo de seleção de feixes para sistemas *Multiple Input Multiple Output* (MIMO) por meio do uso de sinais pilotos, sendo essa uma abordagem mais tradicionalmente adotada pelos órgãos de padronização, como o 3GPP [5]. Tal método, denominado de *baseline*, executa a determinação do melhor feixe para a comunicação por meio de uma série de varreduras ao longo das opções de feixe disponíveis no sistema.

Diferentemente de métodos propostos pautados em IA, que exploram informações provenientes de sensores como câmeras e/ou LiDAR, o método *baseline* dispara sinais de comunicação, denominados de sinais piloto, para sondar o estado do canal de comunicação. Essa operação impõe um consumo adicional de recursos de rádio, caracterizado como *overhead*, durante o processo de gerenciamento de feixes [3]. Em contrapartida, os métodos baseados em IA investigados neste projeto têm como objetivo reduzir ou até mesmo eliminar esse *overhead*, oferecendo alternativas mais eficientes e adaptativas para ambientes complexos e dinâmicos.

2.2 Cenário considerado

Ambos os métodos, o *baseline* e o proposto, serão avaliados em um mesmo cenário de mobilidade urbana, em que há a presença de veículos atuando tanto como receptores quanto como obstáculos. Nesse cenário, é adotada uma frequência de portadora de 60 GHz, garantindo que o sistema opere dentro da faixa de frequência de mmWave. Além disso, considera-se um intervalo de 1 ms entre as medições do canal, em conformidade com uma característica do *software Simulation of Urban Mobility* (SUMO), responsável por simular o movimento dos veículos durante a geração do conjunto de dados de canal. Esse *software* possui como intervalo mínimo de atualização de posição o valor de 1 ms. Dessa forma, considerando a faixa de operação em mmWaves e a granularidade temporal imposta pelo simulador de 1 ms, foi necessário limitar a velocidade máxima dos objetos móveis a 10 km/h, de modo a representar um ambiente de tráfego denso. Essa configuração resulta em um tempo de coerência do canal próximo de 1 ms, o que contribui para reduzir a perda de informação causada por variações excessivas no canal, que em frequências elevadas, como as da faixa de mmWave, tais variações são intensificadas pelos efeitos de *Doppler Spread*.

O conteúdo deste capítulo foi desenvolvido pelos pesquisadores *João Borges*, *Aldebaro Klautau* e *Cristiano B. Both*.

2.3 Metodologia para comparações de desempenho

Para o desenvolvimento do código que implementa o método *baseline* foi usada a linguagem de programação Python, versão 3. O método em si é constituído por três etapas distintas e subsequentes, ilustradas na Figura 1.

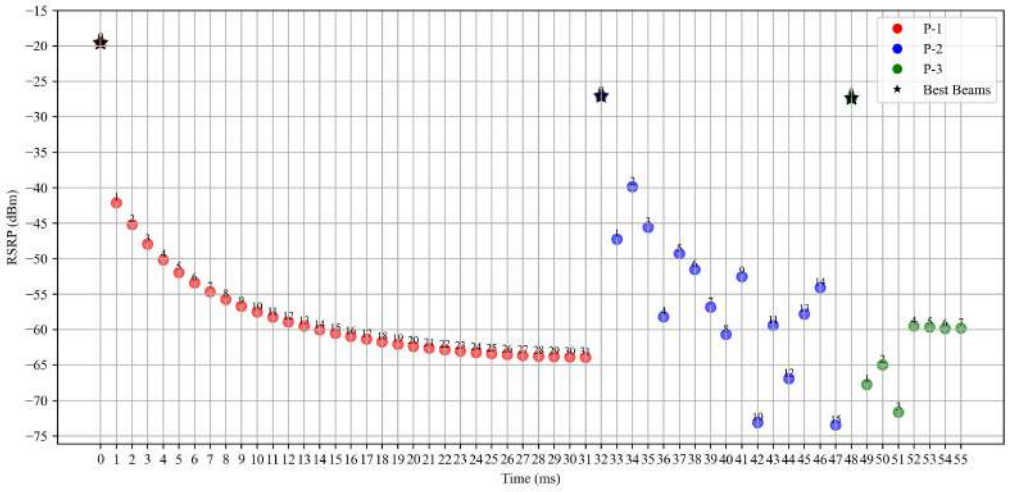


Figura 1: Simulação dos três procedimentos principais definidos pelo 3GPP para o gerenciamento de feixes no 5G: P-1 (varredura inicial por SSBs), P-2 (refinamento de feixe de transmissão via CSI-RS) e P-3 (refinamento de feixe de recepção via CSI-RS).

A primeira etapa, denominada P-1, ocorre no transmissor e utiliza uma categoria de sinais piloto conhecida como *Synchronization Signal Block*s (SSBs). Essa etapa permite estabelecer uma estimativa do direcionamento do feixe mais adequado, além de viabilizar o enlace de acesso inicial entre o transmissor e o receptor. No exemplo considerado, o procedimento P-1 é executado apenas uma vez, com início no instante $t = 0$ ms, seguido por uma varredura sequencial de 32 SSBs de transmissão ao longo de 32 ms consecutivos, sendo 1 ms dedicado a cada SSB. Para cada feixe transmitido, o receptor calcula o valor do RSRP, em dBm, e ao final da varredura, seleciona-se o feixe correspondente ao maior valor de RSRP, que será utilizado como referência na etapa subsequente. Na Figura 1, os pontos correspondentes ao processo P-1 são representados na cor vermelha, com o feixe ótimo sendo indicado por uma estrela. Esse mesmo padrão se repete para os próximos processos, variando apenas a cor dos pontos, com o processo P-2 assumindo a cor azul e o P-3 a cor verde.

Então, logo após o término da etapa anterior, dá-se início à segunda etapa, chamada de P-2, que também ocorre no transmissor e é responsável por uma calibração mais precisa do que a obtida no procedimento P-1. Durante esse procedimento, não são utilizados os SSBs, mas sim outros sinais piloto denominados *Channel State Information Reference Signal*s (CSI-RSs). Com eles, realiza-se uma varredura em uma área mais concentrada, sendo que o receptor retorna um *feedback* contendo informações mais detalhadas, como a potência recebida de cada feixe e outros dados do canal, por meio de um *Channel State Information* (CSI) *feedback*. Na Figura 1, é possível observar o início do procedimento P-2 no instante $t = 32$ ms, imediatamente após o término do P-1, com o transmissor fixando o feixe de recepção em um valor padrão e realizando uma varredura de 16 feixes CSI-RS ao longo do intervalo de $t = 32$ ms a $t = 47$ ms. Assim como no processo P-1, o receptor mede o RSRP para cada feixe e seleciona aquele que apresentar o maior valor como novo feixe de transmissão.

Por último, a terceira etapa, conhecida como P-3, executa um procedimento semelhante ao realizado na etapa P-2, porém desta vez no lado do receptor. O feixe transmissor é mantido fixo, e o receptor realiza uma varredura entre os feixes disponíveis, fazendo uso de CSI-RSs e determinando qual deles possui maior qualidade de sinal. Conforme ilustrado na Figura 1, o procedimento P-3 tem início no instante $t = 48$ ms, sendo seguido pela varredura dos 8 feixes de recepção disponíveis. Ao final do processo, o feixe que apresentar o maior valor de RSRP é selecionado como o melhor feixe de recepção.

Em resumo, a simulação cobre um total de 55 ms, com espaçamento temporal fixo de 1 ms entre transmissões. Os processos P-1, P-2 e P-3 possuem durações de 32 ms (de 0 a 31 ms), 16 ms (de 32 a 47 ms) e 8 ms (de 48 a 55 ms), respectivamente. Nesse intervalo, são avaliadas ao todo 32 transmissões de SSB, 16 transmissões de CSI-RS de transmissão e 8 transmissões de CSI-RS de recepção. Ao final dos procedimentos, os feixes que apresentarem o maior valor de RSRP nos processos P-2 e P-3 são selecionados como os melhores feixes de transmissão e recepção, respectivamente. Vale destacar que a implementação atual não considera eventuais interrupções no processo, as quais podem ocorrer dependendo da qualidade do canal e resultar em custos temporais adicionais.

Em seu estágio atual, o método *baseline* foi testado de forma isolada, utilizando canais de comunicação simples gerados por meio de modelos estocásticos clássicos. No entanto, está previsto seu uso em conjunto com canais provenientes dos *datasets* desenvolvidos pelo projeto, utilizando metodologia multimodal (S013), ou com versões alternativas desses conjuntos de dados, por exemplo, aquelas que empregam uma quantidade maior de cenas por episódio. Essa maior disponibilidade de cenas possibilita o acompanhamento mais detalhado da evolução temporal dos processos mencionados anteriormente (P-1, P-2 e P-3).

3 Expansão de Bases de Dados de Canais para Simulações Suportando Visão Computacional

As redes 5G que operam na faixa de mmWave têm se tornado cada vez mais predominantes, o que torna essencial o estudo de técnicas avançadas, como MIMO e *beamforming*, para mitigar os efeitos do desvanecimento e da elevada suscetibilidade a obstruções. Diante desse cenário, diversas pesquisas vêm adotando métodos baseados em IA com o objetivo de reduzir os impactos causados pelo desalinhamento de feixes, um desafio recorrente em comunicações caracterizadas por alta dinâmica de varredura.

Contudo, para que tais investigações sejam viáveis, é imprescindível dispor de extensos conjuntos de dados. Considerando que campanhas de medição envolvem custos elevados e significativa complexidade operacional, a geração de dados sintéticos surge como uma alternativa crucial, permitindo a simulação de múltiplos cenários de comunicação de forma eficiente e escalável.

A seguir, são apresentados os esforços voltados à criação de bases de dados de canais para simulação de redes 6G, com suporte ao uso de visão computacional.

3.1 Metodologia para geração de dados

Para a geração dos dados, adotou-se a metodologia *Raymobtime*, que integra de forma coordenada três ferramentas computacionais especializadas. O *Wireless Insite* é empregado para simulação avançada da propagação eletromagnética através da técnica de traçado de raios, enquanto o *Blensor* atua na geração sintética de dados de imagem com alto realismo visual. Complementando o processo, o SUMO simula padrões realistas de tráfego veicular, modelando a dinâmica de mobilidade urbana em toda a malha viária do cenário estudado [6].

A metodologia *Raymobtime* opera através de ciclos iterativos de simulação, onde múltiplas execuções sincronizadas do *Wireless Insite* e *Blensor* são coordenadas com os padrões de movimento gerados pelo SUMO [7]. Essa abordagem sistemática viabiliza a construção de um *dataset* capaz de capturar com fidelidade tanto a complexidade do ambiente urbano quanto os intricados fenômenos de propagação em frequências mmWave. Como resultado final do processo, são gerados dois tipos principais de dados sincronizados: os parâmetros do feixe ótimo para cada configuração espacial específica e as correspondentes capturas de imagem da cena [8].

Essa abordagem tem sido amplamente utilizada na literatura para pesquisas envolvendo IA [1, 9, 10, 11], particularmente em problemas de otimização como *beam-selection* e *beam-tracking*. Os *datasets* gerados descrevem tipicamente cenários de comunicação *Vehicle-to-Infrastructure* (V2I) em ambientes urbanos, permitindo diversas estratégias de treinamento. Dentre as aplicações investigadas, destacam-se:

- Seleção de feixes baseada exclusivamente em coordenadas geográficas *Global Positioning System* (GPS);
- Seleção de feixes utilizando apenas dados visuais de câmeras acopladas à antena transmissora;
- Combinação híbrida das duas modalidades, reduzindo, ou até mesmo eliminando, a necessidade de *beam-sweeping*, o que minimiza o overhead de sinalização.

O conteúdo deste capítulo foi desenvolvido pelo pesquisador *Daniel Suzuki* e *Aldebaro Klautau*

Embora a literatura reporte resultados satisfatórios com a metodologia *Raymobtime*, este trabalho propõe aprimoramentos no processo de geração de dados para aumentar a eficiência do sistema, focando principalmente no suporte ao uso de visão computacional. A nova abordagem desenvolvida integra dados de posicionamento do veículo receptor, obtidos via GPS, com parâmetros da câmera, incluindo orientação, distância focal, dimensões do sensor e resolução da imagem, permitindo a estimação precisa da direção da antena receptora ou do veículo alvo no *dataset*.

A transformação de coordenadas GPS para pixels é realizada através de uma projeção 3D-2D, cujo processo envolve três etapas computacionais principais. Primeiramente, calcula-se a distância focal em pixels para cada eixo da imagem mediante as relações dadas por

$$f_x = \frac{f_{mm} \cdot W_{px}}{W_{mm}}, \quad f_y = \frac{f_{mm} \cdot H_{px}}{H_{mm}}, \tag{1}$$

onde f_{mm} representa a distância focal física da lente (em milímetros), (W_{px}, H_{px}) as dimensões da imagem em pixels, e (W_{mm}, H_{mm}) as dimensões físicas do sensor. Nota-se que os valores de f_x e f_y podem diferir devido a possíveis assimetrias no sensor.

Posteriormente, transformam-se as coordenadas globais da antena receptora para o sistema de referência da câmera através da operação dada por

$$\mathbf{P}_{cam} = \mathbf{R} \cdot (\mathbf{Pg}_{ant} - \mathbf{Pg}_{cam}), \tag{2}$$

em que $\mathbf{R}$ representa a matriz de rotação que define a orientação da câmera no sistema global, e $\mathbf{Pg}_{ant}$ e $\mathbf{Pg}_{cam}$ correspondem, respectivamente, às posições tridimensionais da antena e da câmera em coordenadas globais.

Finalmente, a projeção perspectiva no plano imagem é obtida pela operação

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K} \cdot \begin{bmatrix} x_c/z_c \\ y_c/z_c \\ 1 \end{bmatrix}, \quad \text{onde} \quad \mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Aqui, (u, v) denotam as coordenadas em pixels no plano imagem, (x_c, y_c, z_c) são as componentes de $\mathbf{P}_{cam}$, e (c_x, c_y) representam o ponto principal da câmera, tipicamente localizado no centro da imagem. A matriz $\mathbf{K}$, conhecida como matriz de calibração intrínseca, encapsula os parâmetros ópticos do sistema de imageamento.

Esta inovação proporciona informações adicionais para a rede neural, permitindo uma aprendizagem mais precisa da direção ótima de apontamento do feixe (*optimal beam direction*) em relação ao veículo receptor. Conforme ilustrado na Figura 2, o resultado da nova *feature* compreende três dados distintos: (a) a imagem bruta capturada pela câmera, (b) a versão anotada com um marcador específico na antena do veículo, e (c) a imagem processada contendo uma *bounding box* delimitando o veículo receptor.

O método foi sistematicamente aplicado a todos os veículos receptores presentes em cada cena do *dataset*, garantindo assim a consistência dos dados para o treinamento da rede neural. Essa abordagem meticulosa de anotação permite correlacionar diretamente as características visuais com os parâmetros de propagação eletromagnética, estabelecendo uma base robusta para a predição do feixe ótimo. A seguir, são apresentados mais detalhes sobre o *dataset*.

3.2 Dataset

O novo *dataset*, denominado S013, foi gerado mediante a aplicação da metodologia *Raymobtime* no cenário urbano de Rosslyn. A Tabela 1 apresenta os principais parâmetros de

Figura 2: (a) Imagem bruta capturada pela câmera, (b) Imagem com marcador em antena receptora, (c) Imagem com *bounding box* no veículo receptor.

configuração, os quais seguem uma padronização semelhante à adotada nos *datasets* anteriores da série (S008 e S009).

Tabela 1: Características gerais do *dataset*.

Parâmetro	Valor
Frequência Portadora	60GHz
Cenário 3D	Rosslyn
Número de Receptores	10
Número de Episódios	2000
Cenas por episódio	1
Tempo entre Episódios	30s
Número de Canais Válidos	20k

Além dos dados convencionais de GPS e imagens já disponíveis nos *datasets* padrão do *Raymobtime*, a nova versão apresentada, S013, incorpora novas *features*. Primeiramente, inclui imagens adicionais, conforme mencionado anteriormente, que marcam explicitamente a localização da antena receptora e do veículo correspondente, oferecendo maior flexibilidade para os usuários que desejam empregar técnicas baseadas em visão computacional. Adicionalmente, complementando os dados de GPS, o *dataset* agora fornece as coordenadas locais (P_{cam}) de cada veículo em relação a cada câmera da antena transmissora.

Para enriquecer ainda mais a representação do ambiente, a configuração de câmeras foi expandida de 3 para 6 unidades, organizadas em pares cobrindo cada direção principal. Esta ampliação da cobertura angular proporciona uma visão mais abrangente do cenário, habilitando novas possibilidades de pesquisa em percepção visual e estereoscopia.

Esta nova base de dados compreende aproximadamente 20 mil amostras, uma quantidade adequada para o treinamento de redes neurais, superando as versões anteriores do *Raymobtime* tanto em volume quanto em estrutura de dados — aproximadamente o dobro do que foi disponibilizado nas versões S008 e S009. A disponibilização do S013 permite uma comparação direta com os resultados obtidos nesses conjuntos anteriores, facilitando a avaliação de novas técnicas de aprendizado de máquina em condições controladas.

4 Busca de Arquiteturas Neurais Aplicada à Seleção de Feixes

O problema de seleção de feixes (*beam selection*) em enlaces mmWave consiste em identificar, de forma eficiente, o subconjunto *top-K* de pares transmissor–receptor que maximize a potência recebida, evitando a busca exaustiva em todo o *codebook* de feixes. Neste trabalho, propõe-se o uso de *Neural Architecture Search*s (NASs), implementadas com a biblioteca `Optuna` e o *sampler Tree-structured Parzen Estimator* (TPE), para buscar automaticamente a melhor arquitetura de rede neural capaz de prever o conjunto *top-K* de feixes. A otimização da arquitetura explora dinamicamente hiperparâmetros como o número de camadas, a quantidade de neurônios por camada, as funções de ativação e as taxas de aprendizado, visando minimizar a função de perda de classificação sobre o conjunto de treino, com validação em dados independentes [1].

O conjunto de dados *Raymobtime* está dividido em dois cenários de tráfego distintos: S008 (tráfego regular) e S009 (horário de pico — *rush hour*). O cenário S008 contém 11.194 amostras, sendo 6.482 do tipo *Line-of-Sight* (LoS) e 4.712 do tipo *Non-Line-of-Sight* (NLoS), e é utilizado para o treinamento dos modelos. Já o cenário S009 compreende 9.638 amostras, das quais 1.473 são LoS e 8.165 são NLoS, sendo empregado para fins de validação.

4.1 Trabalhos relacionados

Salehi et al. [1] propuseram um *framework* de fusão profunda de múltiplas modalidades, denominado *Fusion-based Deep Learning* (F-DL), com o objetivo de realizar a seleção *top-K* de feixes em sistemas de comunicação mmWave. A abordagem é baseada na ideia de explorar diferentes fontes de informação — ou modalidades — como coordenadas GPS, imagens e dados de LiDAR, a fim de melhorar a robustez e a acurácia da predição de feixes. Cada modalidade é inicialmente processada por uma rede neural unimodal: *conv1D* é utilizada para as coordenadas GPS, enquanto blocos do tipo ResNet-like são empregados para processar tanto as imagens quanto os dados de LiDAR. O resultado de cada rede unimodal é um vetor latente (*embedding*) de dimensão d, representando as características extraídas da respectiva entrada. Esses vetores latentes são então concatenados em um único vetor, que é processado por uma rede de fusão responsável por produzir, via *softmax*, uma pontuação (*score*) para cada par de feixes possíveis. O modelo é treinado *offline* utilizando a função de entropia cruzada, visando maximizar a probabilidade de seleção dos feixes corretos. Durante a inferência, o sistema é projetado para operar de forma distribuída entre o veículo e os *Mobile Edge Computing*s (MECs), com o objetivo de reduzir a latência de comunicação e acelerar a tomada de decisão em tempo real.

Os resultados reportados por Salehi et al. no conjunto S009 demonstram ganhos significativos ao combinar modalidades. A Tabela 2 resume a acurácia *top*-1 e *top*-10 para cada configuração de entrada.

4.2 Metodologia

A definição das melhores arquiteturas foi conduzida por meio da técnica de NAS, utilizando a biblioteca `Optuna`[12], enquanto a criação e o treinamento das redes neurais foram realizados com a biblioteca `Keras`, tendo o `TensorFlow` como *backend* [13]. Cada arquitetura foi treinada

O conteúdo deste capítulo foi desenvolvido pelos pesquisadores *Matheus F. Silva e Felipe A. P. de Figueiredo*.

Tabela 2: Acurácia *top*-1 e *top*-10 no conjunto S009. *Tabela adaptada de Salehi et al.*[1]

Modalidades	***Top*-1 Accuracy**	***Top*-10 Accuracy**
GPS	12.32 %	77.93 %
Imagem	12.39 %	71.65 %
LiDAR	46.23 %	89.95 %
GPS + Imagem	25.76 %	86.29 %
GPS + LiDAR	55.42 %	91.41 %
Imagem + LiDAR	54.52 %	91.23 %
GPS + Imagem + LiDAR	56.22 %	91.11 %

por, no máximo, 50 épocas. Para guiar a busca pelas melhores configurações, empregou-se o algoritmo de otimização Bayesiana TPE, que explora o espaço de busca com base em distribuições condicionais estimadas a partir do desempenho dos modelos avaliados.

Foram investigadas arquiteturas baseadas em camadas `Conv1D`, `Conv2D` e `Conv3D`, combinadas com diferentes estratégias de pré-processamento de dados e variações nos principais hiperparâmetros da rede.

O espaço de busca abrangeu, entre outros hiperparâmetros: a normalização dos dados de entrada, o tipo de camada principal, o número de unidades ou filtros, as funções de ativação, a profundidade da rede, a inclusão de camadas densas adicionais e seus respectivos parâmetros. Além disso, foram aplicadas estratégias complementares de otimização, incluindo a normalização das entradas, o uso de *early stopping* e ajustes dinâmicos da taxa de aprendizado ao longo do treinamento. O processo de otimização da arquitetura baseada em `Conv1D` está descrito no Algoritmo 1.

As Figuras 3a e 3b apresentam, de forma complementar, a exploração do hiperparâmetro `learningRate` ao longo do processo de busca conduzido com a biblioteca `Optuna`. A Figura 3a mostra um histograma dos valores de `learningRate` testados em todos os *trials*, sobreposto a uma curva de densidade estimada, o *Kernel Density Estimation* (KDE). A caixa de texto exibe as principais estatísticas: média (3.73×10^{-5}), desvio padrão (5.40×10^{-6}) e mediana (3.66×10^{-5}). As linhas verticais tracejada (vermelha) e contínua (verde) representam, respectivamente, a média e a mediana, indicando que a maior concentração de valores testados se situou em torno de 3.3×10^{-5} a 3.9×10^{-5}.

Já a Figura 3b compara, lado a lado, os histogramas de todos os *trials* (barras em cinza) com aqueles referentes aos 20% de melhores *trials* (barras em verde). A linha tracejada vermelha assinala o intervalo interquartil, isto é, a faixa compreendida entre os percentis 25% e 75%, que abrange os 50% centrais das observações. A linha pontilhada azul indica o intervalo entre os percentis 5% e 95%, cobrindo 90% dos dados ao excluir os 5% menores e os 5% maiores valores. Já a linha preta contínua marca a mediana dos *trials* de melhor desempenho. Observa-se que, nos experimentos mais eficazes, o valor de `learningRate` concentrou-se em uma faixa ainda mais restrita, aproximadamente entre 3.33×10^{-5} e 3.69×10^{-5}, o que sugere essa região como particularmente promissora para futuras otimizações.

4.3 Trabalhos Futuros

A pesquisa encontra-se em andamento e, nas próximas etapas, a otimização das arquiteturas avaliadas (`Conv1D`, `Conv2D` e `Conv3D`) terá como foco o ajuste de hiperparâmetros para equilibrar a acurácia nas métricas *top*-1 e *top*-10, a latência de inferência e o tamanho do modelo. Serão

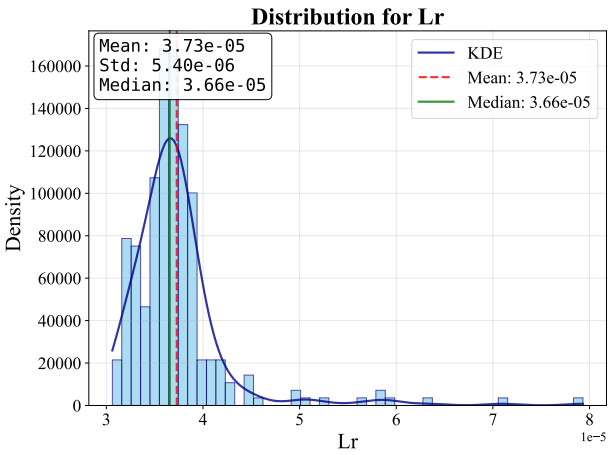
Entrada: Conjunto de dados $\mathcal{D}$; número máximo de trials T; número de épocas E
Saída: Melhor conjunto de hiperparâmetros, Θ^*
Divida $\mathcal{D}$ em $\mathcal{D}_{\text{train}}$ e $\mathcal{D}_{\text{val}}$;
for Trial $\leftarrow 1$ **to** T **do**

- **Normalização dos dados LiDAR: Escolha** $\theta_{\text{normLiDAR}} \in \{none, minmax, standard\}$
- **Normalização de coordenadas: Escolha** $\theta_{\text{normCoord}} \in \{minmax, standard\}$
- **Canais LiDAR: Escolha** $\theta_{\text{canaisLiDAR}} \in \{4, 14\}$
- **Número de camadas Conv1D: Escolha** $\theta_{\text{layersConv1D}} \in [3, 6]$
- **Filtros Conv1D: Escolha** $\theta_{\text{filtrosConv1D}} \in \{32, 64, \ldots, 512\}$
- **Tamanho do kernel Conv1D: Escolha** $\theta_{\text{kernelConv1D}} \in [1, 5]$
- **Tamanho do pool Conv1D: Escolha** $\theta_{\text{poolConv1D}} \in [1, P_{\max}]$
- **Camadas densas adicionais: Escolha** $\theta_{\text{denseCamadas}} \in [0, 3]$
- **Unidades por camada densa: Escolha** $\theta_{\text{unidadesDense}} \in \{50, 100, \ldots, 500\}$
- **Tamanho de batch: Escolha** $\theta_{\text{batchSize}} \in \{64, 128, 256\}$
- **Função de ativação: Escolha** $\theta_{\text{ativacao}} \in \{relu, tanh, sigmoid, swish\}$
- **Regularização: Escolha** $\theta_{\text{regularizacao}} \in \{none, l1, l2, l1l2\}$
- **Otimizador: Escolha** $\theta_{\text{otimizador}} \in \{AdamW, Lion, RMSprop\}$
- **Scaler de entrada: Escolha** $\theta_{\text{scalerEntrada}} \in \{StandardScaler, MinMaxScaler\}$
- **Learning rate: Escolha** $\theta_{\text{learningRate}} \in [10^{-5}, 10^{-2}]$
- **Inicializador de pesos: Escolha** $\theta_{\text{inicializador}} \in \{Zeros, Ones, Const., RN, RU, TN, GN, GU, HN, HU, LN, LU, Id., Orthogonal, VS\}$
- **Batch normalization: Escolha** $\theta_{\text{batchNorm}} \in \{sim, não\}$
- **Dropout: Escolha** $\theta_{\text{dropoutCamadas}} \in [0, 3]$ *e* $\theta_{\text{dropoutTaxa}} \in \{0.0, 0.1, \ldots, 0.5\}$

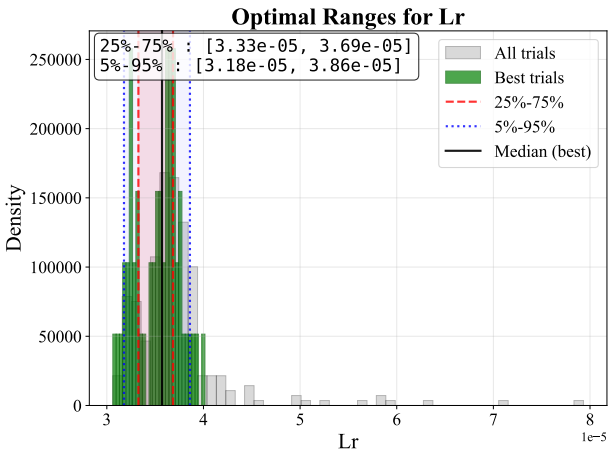
end
return Θ^*

Algorithm 1: Processo de otimização para a arquitetura utilizando `Conv1D`

empregadas técnicas como poda estrutural, quantização de pesos e busca por variantes de arquiteturas mais leves, explorando diferentes combinações de profundidade de rede, largura de canais e tipos híbridos de convolução. Além disso, outras arquiteturas promissoras, como os *transformers*, também serão investigadas. Cada configuração será avaliada por meio de validação cruzada *k-fold* no cenário S008 e validação independente no cenário S009, com registro de curvas de aprendizado, tempo de treinamento e consumo de memória. Os resultados obtidos por simulação fornecerão subsídios importantes para o refinamento de diretrizes de projeto e para orientar futuras implementações em ambientes reais.



(a) Distribuição numérica do hiperparâmetro `learningRate` obtida no teste.



(b) Intervalos do hiperparâmetro `learningRate` observados no teste.

Figura 3: Distribuições e tendências do hiperparâmetro `learningRate` obtidas durante um teste.

5 Visão Computacional e Dados Multimodais para Seleção de Feixes

A 6G impõe novos paradigmas para a infraestrutura de comunicação, com ênfase em conectividade ubíqua, latência ultrabaixa (< 1 ms), taxas de transmissão da ordem de Terabits por segundo (Tbps) e suporte nativo a IA distribuída. Em cenários de mobilidade elevada e ambientes urbanos densos, essas exigências se traduzem na necessidade de enlaces altamente direcionalizados, adaptativos e resilientes a bloqueios e interferências.

A operação em mmWave e Terahertz (THz), previstas como pilares do espectro 6G, oferece grande capacidade espectral, mas introduz desafios relevantes, como alta perda de propagação, baixa difração e sensibilidade a oclusões físicas. Para mitigar tais efeitos, são empregados arranjos massivos de antenas com formação de feixe (*beamforming*), exigindo mecanismos eficientes de seleção de feixes (*beam selection*) entre transmissor e receptor. Contudo, abordagens baseadas em varredura exaustiva, como nos padrões IEEE 802.11ad e 5G-*New Radio* (5G-NR), impõem tempos de latência incompatíveis com os requisitos operacionais de aplicações veiculares em 6G, especialmente considerando a curta duração dos períodos de contato (*link duration*) entre um veículo em movimento e a infraestrutura de acesso.

Diante dessas limitações, a exploração de dados contextuais exógenos à camada física, oriundos de sensores embarcados e de infraestrutura, como GPS, câmeras RGB e sensores LiDAR, surge como uma estratégia promissora para restringir o espaço de busca dos feixes candidatos e acelerar o processo de alinhamento direcional. A utilização de técnicas de visão computacional, combinadas com modelos de aprendizado profundo e inferência distribuída na borda da rede (*edge intelligence*), possibilita a estimativa probabilística dos feixes mais promissores (*top-K beam pairs*), reduzindo significativamente a complexidade computacional e o tempo de resposta do sistema.

Nesse contexto, a formulação de arquiteturas de fusão multimodal robustas, capazes de operar em tempo real com múltiplas modalidades sensoriais, e a incorporação de critérios de otimização adaptativos para seleção do espaço de busca, constituem elementos centrais para viabilizar a comunicação em redes 6G com requisitos de confiabilidade, eficiência espectral e latência extrema.

5.1 Trabalhos relacionados

Salehi et al. [1] propõem um método de seleção de feixes para comunicação em redes veiculares usando mmWave. A proposta utiliza dados sensoriais multimodais (GPS, câmera e LiDAR) combinados por meio de uma arquitetura de rede neural, denominada F-DL, para prever os top-K pares de feixes mais prováveis, reduzindo o espaço de busca necessário para estabelecer o enlace e, consequentemente, reduzindo o tempo de varredura exigido pelas normas 5G-NR e IEEE 802.11ad. A arquitetura conta com redes neurais específicas por sensor, que extraem *embeddings* locais no veículo (para GPS e LiDAR), enviados via canal sub-6 GHz para o MEC. No MEC, esses *embeddings* são combinados com as informações visuais processadas na estação base, permitindo a predição do feixe ideal. Um algoritmo adicional seleciona dinamicamente o valor de K, balanceando a probabilidade de acerto e o tempo de varredura. Os experimentos, realizados com o conjunto de dados simulado *Raymobtime* e o conjunto de dados real NEU,

O conteúdo deste capítulo foi desenvolvido pelos pesquisadores *Wesley L. Passos* e *José F. Rezende*.

mostram que a abordagem reduz em até 96% o tempo de seleção de feixes, mantendo até 97,95% do *throughput*. A técnica também melhora a acurácia na predição dos melhores feixes em até 22% em relação ao estado da arte. Embora a abordagem proposta por Salehi et al. [1] traga ganhos expressivos em tempo e acurácia na seleção de feixes, sua aplicação prática enfrenta limitações. O método depende de sensores multimodais integrados e sincronizados, o que pode ser inviável em cenários com infraestrutura limitada. Além disso, a robustez do sistema em condições adversas não é suficientemente avaliada, e o modelo apresenta alta complexidade de treinamento. Portanto, apesar do potencial, a solução ainda exige adaptações para ser aplicável em contextos reais e diversos.

5.2 Trabalhos desenvolvidos

Esta seção apresenta as atividades realizadas no escopo do projeto até o momento, com foco na reprodução e validação de abordagens de referência para comunicação veicular em redes 6G. Primeiramente, buscou-se estabelecer um *baseline* experimental a partir da reprodução do trabalho de Salehi et al. [1], com as devidas adaptações técnicas necessárias. Em seguida, foram conduzidos experimentos preliminares voltados à avaliação de técnicas de segmentação semântica, com vistas a melhorar a percepção visual para a tarefa de seleção de feixes.

5.2.1 Estabelecimento do *baseline* experimental

Para estabelecer um *baseline* de referência, buscou-se inicialmente reproduzir os resultados apresentados por Salehi et al. [1]. Como descrito na Seção 5.1, o trabalho propõe uma abordagem baseada em aprendizado profundo para acelerar a seleção de feixes em comunicações mmWave a partir de dados multimodais (GPS, câmera e LiDAR). Entretanto, ao tentar utilizar o código originalmente disponibilizado pelos autores, identificou-se que a versão do *framework* de aprendizado profundo utilizada (`TensorFlow`) estava desatualizada e incompatível com os ambientes atuais de desenvolvimento.

Diante disso, optou-se por adaptar o código para uma versão mais recente do `TensorFlow`, o que exigiu uma refatoração de diversas partes do *pipeline*, incluindo funções de modelagem, treinamento e inferência. Essa etapa foi importante para assegurar a execução correta dos experimentos e a replicabilidade dos resultados descritos no artigo original. O código resultante dessa adaptação foi disponibilizado publicamente em repositório GitHub [14], contribuindo para a reprodutibilidade e disseminação dos resultados no contexto do projeto Brasil 6G.

Após a etapa de refatoração, foram conduzidos experimentos de avaliação da arquitetura de fusão em diferentes combinações de modalidades unimodais, conforme descrito por Salehi et al. [1], utilizando o conjunto de dados *Raymobtime*. Os modelos foram treinados sobre o subconjunto S008 e avaliados no subconjunto S009, replicando o protocolo experimental original.

A Tabela 3 apresenta os resultados obtidos com o código refatorado em comparação com os valores reportados no artigo. Observa-se que os desempenhos reproduzidos foram, em geral, compatíveis com os originalmente reportados, especialmente em termos de acurácia na seleção de feixes (*top-K accuracy*). Pequenas variações observadas podem ser atribuídas a diferenças nas versões das bibliotecas utilizadas, na inicialização dos pesos ou na aleatoriedade inerente ao treinamento dos modelos.

Tabela 3: Resultados de *top*-1 e *top*-10 *accuracy* para diferentes combinações de modalidades sensoriais, comparando os valores originalmente reportados por Salehi et al. [1] com os obtidos a partir da reprodução e refatoração do código.

Modalidades	*Top*-1		*Top*-10	
	Reportado	Reproduzido	Reportado	Reproduzido
Coord.	12,32%	12,32%	77,93%	77,93%
Imagem	12,39%	12,45%	71,65%	72,32%
LiDAR	46,23%	51,65%	89,95%	90,59%
Coord. + Imagem	25,76%	23,25%	86,29%	84,20%
Coord. + LiDAR	55,42%	52,75%	91,41%	90,59%
Imagem + LiDAR	54,52%	54,60%	91,23%	90,70%
Coord. + Imagem + LiDAR	56,22%	54,51%	91,11%	90,37%

5.2.2 Resultados iniciais de segmentação utilizando YOLOv11

Uma primeira direção promissora consiste em aumentar a robustez da percepção visual por meio do uso de técnicas de segmentação semântica [15], que permitem identificar com maior precisão diferentes classes de objetos, como veículos, pedestres e obstáculos estáticos, e integrar essas informações ao processo de seleção de feixes. Como experimento preliminar, foi utilizado o modelo *You Only Look Once* (YOLO)v11 [16] para detectar, nas imagens, a posição dos transmissores e receptores. Os resultados estão apresentados na Figura 4. Observa-se que o modelo foi capaz de identificar corretamente os veículos presentes nas cenas. No entanto, objetos parcialmente visíveis (recortados nas bordas da imagem) não foram detectados, o que indica uma limitação do modelo em sua configuração atual. Cabe destacar que, até o momento, nenhuma etapa de *fine-tunning* foi aplicada, o que sugere que há margem para melhoria significativa nos resultados com o ajuste fino dos parâmetros do modelo em dados específicos do domínio.

5.3 Direções futuras

Diante dos avanços apresentados, há diversas oportunidades para aprimorar a abordagem proposta e ampliar sua aplicabilidade em cenários mais realistas e complexos. Uma primeira direção promissora consiste em aumentar a robustez da percepção visual. Para isso, sugere-se o uso de técnicas de segmentação semântica [15] que permitam identificar com maior precisão diferentes classes de objetos (como veículos, pedestres e obstáculos estáticos), integrando essas informações ao processo de seleção de feixes, ou ainda na geração de bases de dados sintéticas, via simulações baseadas em *raytracing*.

Além disso, a adoção de modelos multimodais mais avançados, como arquiteturas baseadas em *transformers* [17] ou no *Contrastive Language-Image Pre-training* (CLIP) [18], pode ampliar a expressividade das representações visuais e contextuais da cena. O CLIP, em particular, permite extrair *embeddings* visuais semânticos mais ricos e generalizáveis, que podem substituir redes convolucionais tradicionais nos *pipelines* de fusão multimodal, capturando interações complexas entre veículos, pedestres e infraestrutura. Sua capacidade de integrar descrições textuais gera uma fonte adicional de contexto, útil para ajustar dinamicamente o espaço de busca (*top*-K) conforme a complexidade da cena. Além disso, o CLIP favorece a transferência

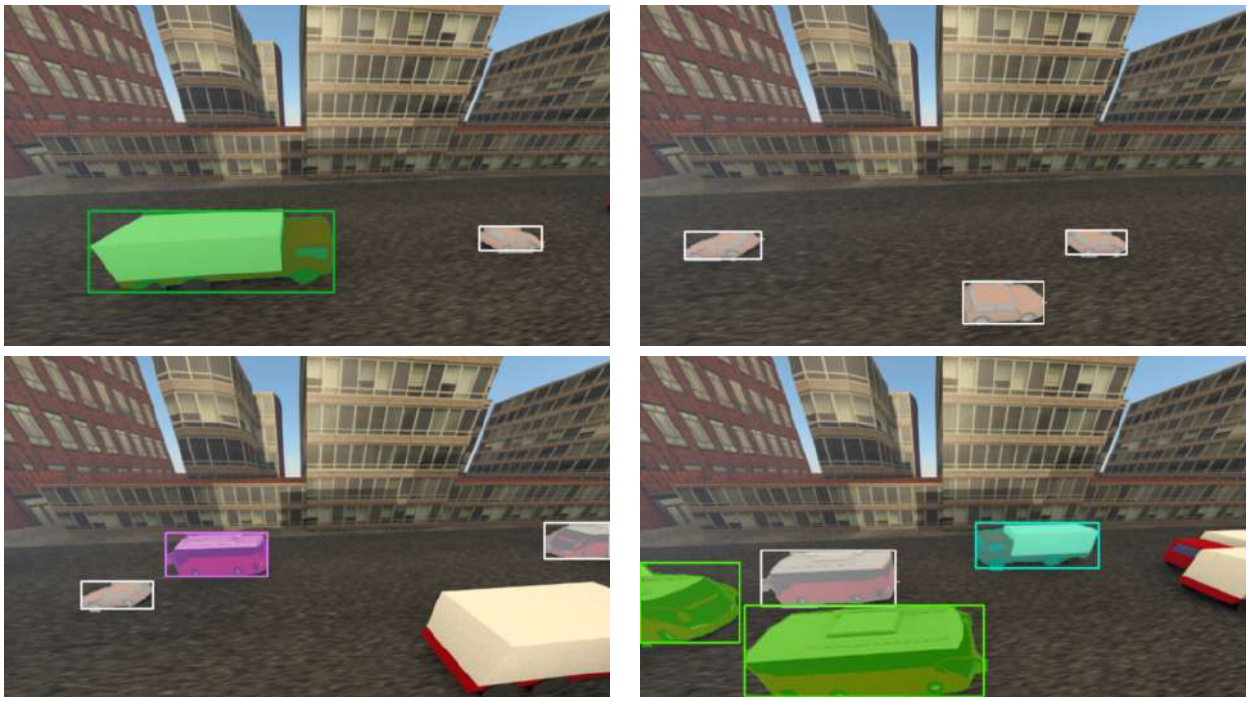


Figura 4: Resultados de segmentação de veículos em imagens simuladas do conjunto *Raymobtime*, utilizando YOLOv11. As caixas delimitadoras (*bounding boxes*) indicam as detecções realizadas.

zero-shot para ambientes urbanos distintos, sem necessidade de retreinamento intensivo. Por fim, a combinação de múltiplas câmeras com diferentes ângulos e técnicas de estereoscopia para estimativa de profundidade e percepção 3D surge como uma estratégia para enriquecer a modelagem do ambiente e ampliar a robustez da percepção visual.

Uma segunda linha de pesquisa relevante refere-se à detecção dinâmica de oclusões e à modelagem de reflexões de sinal. Em vez de operar com imagens estáticas, a análise de sequências de vídeo pode permitir a previsão do movimento de objetos e obstáculos, aumentando a antecipação de perdas de LoS e favorecendo estratégias proativas de seleção de feixes. Complementarmente, a aprendizagem de padrões típicos de reflexão em superfícies metálicas, como as de veículos, por meio de técnicas de visão computacional, pode contribuir para estimativas mais precisas da propagação do sinal em cenários urbanos complexos.

Por fim, destaca-se a importância de validar essas abordagens em outros conjuntos de dados, como o DeepSense 6G [19, 20] (voltado para percepção sensorial e modelagem de ambiente em veículos autônomos), o que possibilitará avaliar a generalização dos modelos propostos frente a diferentes condições ambientais, topologias e dinâmicas de tráfego.

6 Conclusão

Diante dos desafios impostos pelas comunicações em mmWaves, torna-se essencial explorar soluções inovadoras que garantam a eficiência e a confiabilidade do *beam management* em redes 5G e futuras redes 6G. As abordagens baseadas em IA, especialmente aquelas que integram visão computacional, despontam como alternativas promissoras para superar as limitações dos métodos convencionais.

Os resultados alcançados até o momento na execução da Atividade 3.1 do Projeto Brasil 6G incluem novas bases de dados, novas arquiteturas para uso de redes neurais em *beam management* e exploração de novos métodos para uso de visão computacional na redução do *overhead* de sistemas 6G.

Ao alinhar-se com os casos de uso definidos pelo 3GPP, esta atividade reforça a relevância da visão computacional e da aplicação de IA no contexto de padronização e evolução tecnológica das comunicações móveis. A integração de sensores adicionais e algoritmos de aprendizado abre caminho para uma nova geração de sistemas de comunicação mais adaptativos e inteligentes, em especial com a ênfase dada ao *Integrated Sensing and Communications* (ISAC) no 6G.

Os resultados esperados têm o potencial de influenciar positivamente as futuras Releases do 3GPP, contribuindo para o desenvolvimento de soluções mais robustas, de baixa latência e com melhor aproveitamento espectral. Em última instância, o projeto busca avançar o estado da arte no uso da visão computacional em 6G, em tópicos como *beam management*, consolidando o papel da visão computacional e da IA como elementos-chaves na construção das redes 6G.

Referências

[1] B. Salehi, G. Reus-Muns, D. Roy, Z. Wang, T. Jian, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7639–7655, 2022.

[2] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, 2016.

[3] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2019.

[4] M. Zecchin, M. B. Mashhadi, M. Jankowski, D. Gündüz, M. Kountouris, and D. Gesbert, "LIDAR and Position-Aided mmWave Beam Selection With Non-Local CNNs and Curriculum Training," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2979–2990, 2022.

[5] 3rd Generation Partnership Project (3GPP), "Study on New Radio Access Technology: Physical Layer Aspects (Release 14)," 3rd Generation Partnership Project (3GPP), 3GPP Technical Report TR 38.802 V14.2.0, September 2017, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3066.

[6] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *2018 Information Theory and Applications Workshop (ITA)*, 2018, pp. 1–9.

[7] A. Nascimento, W. Frazão, A. Oliveira, D. Gomes, and A. Klautau, "Multimodal dataset for machine learning applied to telecommunications," *SBrT, Santa Catarina, Brazil*, pp. 13–16, 2020.

[8] D. Suzuki, A. Oliveira, L. Gonçalves, I. Correa, A. Klautau, S. Lins, and P. Batista, "Ray-Tracing MIMO Channel Dataset for Machine Learning Applied to V2V Communication," in *2022 IEEE Latin-American Conference on Communications (LATINCOM)*. IEEE, 2022, pp. 1–6.

[9] A. Oliveira, D. Suzuki, S. Bastos, I. Correa, and A. Klautau, "Machine learning-based mmwave mimo beam tracking in V2I scenarios: Algorithms and datasets," in *2024 IEEE Latin-American Conference on Communications (LATINCOM)*. IEEE, 2024, pp. 1–5.

[10] E. Chatzoglou and S. K. Goudos, "Beam-selection for 5G/b5G networks using machine learning: A comparative study," *Sensors*, vol. 23, no. 6, p. 2967, 2023.

[11] S. Bastos, A. Oliveira, D. Suzuki, L. Gonçalves, I. Sousa, and A. Klautau, "Generation of 5G/6G Wireless Channels Using Raymobtime with Sionna's Ray-Tracing," *XLI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, 2023.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *arXiv preprint arXiv:1907.10902*, 2019. [Online]. Available: https://arxiv.org/abs/1907.10902

[13] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," *arXiv preprint arXiv:1605.08695*, 2016. [Online]. Available: https://arxiv.org/abs/1605.08695

[14] W. L. Passos, "TVT_beam_selection," 2025. [Online]. Available: https://github.com/wesleylp/TVT_beam_selection

[15] D. Guo, Y. Pei, K. Zheng, H. Yu, Y. Lu, and S. Wang, "Degraded image semantic segmentation with dense-gram networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 782–795, 2020.

[16] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 1–11.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[19] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-position multimodal beam prediction using real millimeter wave datasets," in *IEEE Wireless Communications and Networking Conference*, 2022, pp. 2727–2731.

[20] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 122–128, 2023.